# Uso de la Web Semántica en la extracción de datos para la evaluación de sitios de gobierno electrónico

**Aristides Dasso*, Ana Funes***
***Universidad Nacional de San Luis**
**Argentina**

## Resumen

En la línea de investigación aquí presentada, nos ocupamos de la aplicación y propuesta de técnicas de extracción de datos de la web para la evaluación de sitios de gobierno electrónico (e-gov). Para llevar a cabo la evaluación de un sitio de e-gov, nuestros modelos hacen uso de los valores observados en el sitio para un conjunto de características pre-establecidas. Es por ello que necesitamos de técnicas automáticas o semi-automáticas que aprovechando tanto los estándares establecidos por los lenguajes HTML y XML, como aquellos brindados por las tecnologías de la Web Semántica nos provean con los datos necesarios para llevar adelante dichas evaluaciones.

***Palabras clave:*** *Sitios de Gobierno electrónico. Modelos de evaluación. Web Semántica.*

## Contexto

Este trabajo de investigación se encuentra enmarcado dentro del Proyecto de Incentivos código 22/F822: "Ingeniería de Software: Conceptos, Métodos y Herramientas en un Contexto de Ingeniería de Software en Evolución", de la Universidad Nacional de San Luis, en la línea "Métodos Formales y Prototipos Evolutivos", del mismo. Dentro del contexto de desarrollo de métodos y herramientas, esta investigación tiene como objetivo el concretar la construcción de una herramienta de software que sirva para la recolección de los datos necesarios para alimentar modelos de evaluación de servicios provistos en sitios de e-gov.

## Introducción

En trabajos previos hemos desarrollado modelos para evaluación de servicios de e-gov. Dichos modelos han sido creados aplicando el método Logic Score of Preferences (LSP) el cual hace uso de una Lógica Continua ([Duj07], [Duj96], [Duj97], [DB97]). Con el objeto de determinar cuáles debían ser los servicios presentes, a brindar al ciudadano en las diferentes áreas que le son propias al estado, en los sitios de gobierno electrónico, comenzamos tomando como punto de partida los indicadores planteados por la Comisión de Gobierno Electrónico de la Unión Europea (eEurope) [EC02], para luego ir extendiéndolo hasta producir nuestro modelo final [Cas10].

En el presente trabajo apuntamos a obtener los datos necesarios para nuestro modelo de evaluación de sitios de e-gov, con el objeto de llevar adelante sus respectivas evaluaciones, haciendo uso de técnicas que saquen ventaja de tecnologías de la Web Semántica [LHL01], de manera de hacer que el proceso de captura de estos datos sea menos dependiente de la intervención humana. Es así que las alternativas a analizar consideran no sólo la estructura actual de la Web sino también las tecnologías inherentes a la Web Semántica que puedan ser relevantes a nuestra tarea [DGC04].

## Resultados y Objetivos

Dado que nuestro objetivo es evaluar los servicios ofrecidos en los sitios de e-gov, esto hace inevitable la navegación de los mismos para obtener la información relevante. Normalmente, esta navegación es realizada en forma manual por humanos, consumiendo una gran cantidad de tiempo. En consecuencia, apuntamos a que dicha navegación sea llevada a cabo por una herramienta de software que tenga incorporada el conocimiento necesario para poder reconocer las distintas características de los sitios (requerimientos) que tengan relevancia con respecto al modelo de evaluación que está siendo usado para la evaluación.

En la Figura 1 mostramos una posible estructura de la Web Semántica. La idea es que la aplicación obtenga datos para la evaluación en distintos niveles. El problema mayor es que en la actualidad, la mayoría de los sitios se sitúan en los niveles más bajos (HTML, XML); es decir que las páginas web de dónde extraer la información no contienen información semántica. Sin embargo, hay sitios que adhiriendo a una ontología, facilitan la tarea de obtención de la información relevante. Es por eso que pensamos que las características propias de la Web Semántica pueden ayudarnos a que la evaluación sea realizada de manera más automática y precisa.

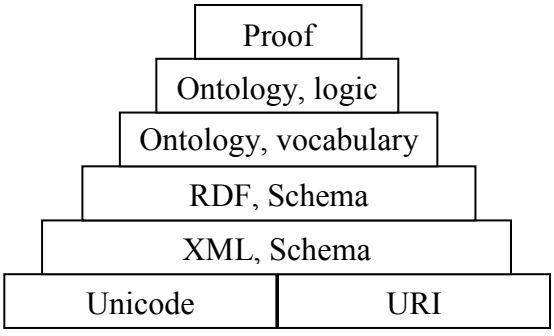


**Figura 1. Estructura de la Web Semántica.**

Resulta claro que extraer información de sitios que se encuentran en los niveles más bajos, y que son los más comunes en la actualidad, es mucho más difícil y requiere de tecnologías propias del campo de la extracción automática de conocimiento o se ve limitada a la búsqueda de palabras clave con pesos, pero planas e inconexas, lo que no permite reconocer ni solicitar significados más elaborados. En este nivel, por ejemplo, trabajan distintos buscadores empleando 'crawlers'.

Una posible aproximación, que contempla el uso de tecnologías de la Web Semántica, para atacar nuestro problema de recolección de datos para evaluación podría ser hecha por medio de una aplicación dirigida por una ontología. De esta forma, esta aplicación podría extraer la información pertinente a cada uno de los ítems de los modelos de evaluación y luego enviar dicha información a la herramienta que realiza la evaluación.

De esto se desprende que el comportamiento del software tendrá que incorporar indudablemente la ontología del sitio a evaluar junto con el correspondiente modelo de evaluación, al que llamaremos ontología de referencia, debiendo construir para esto una correspondencia entre los ítems de los árboles de requerimientos del modelo de evaluación y la ontología del sitio. Cabe aclarar que cuando hablamos de la ontología del sitio, nos referimos sólo a la parte de la ontología que describe la información semántica presente en el sitio en cuestión.

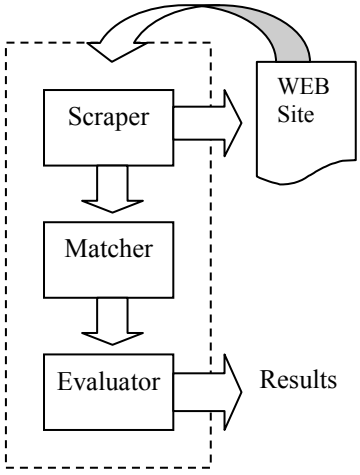


**Figura 2. Estructura general de la aplicación.**

En la Figura 2 podemos ver la estructura general de la aplicación. Ahí pueden verse los tres principales componentes: el "Scraper", el "Matcher" y el "Evaluator".

La tarea principal del Scraper es obtener la ontología del sitio de manera que el Matcher pueda realizar su tarea. Mientras obtiene la ontología del sitio, el Scraper no realizará ningún preprocesamiento de los datos, por lo que esta etapa es totalmente automática. La figura 3 muestra la estructura general del Scraper.

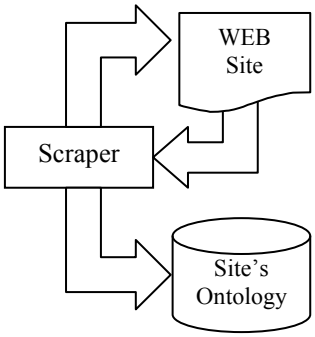


**Figure 3. Estructura del Scraper**

El Matcher recibe los datos (ontología del sitio) del Scraper, y empleando la Ontología de Referencia (dada por el árbol de requerimientos construido en el modelo de evaluación) y un Thesaurus lleva a cabo un proceso de alineamiento de las ontologías.

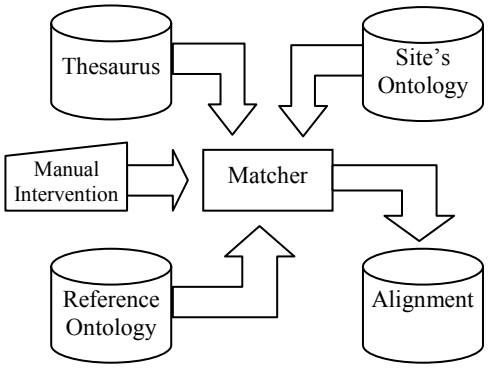


**Figura 4. Estructura del Matcher**

Las clases así como algunas de las relaciones son comparadas usando el Thesaurus. La figura 4 muestra una estructura general del Matcher.

Sin embargo, el proceso no es completamente automático y se requiere la intervención humana, especialmente para el matching entre las relaciones o cuando las clases tienen diferencias estructurales importantes. La herramienta debería asistir al operador identificando las diferencias entre ambas ontologías.

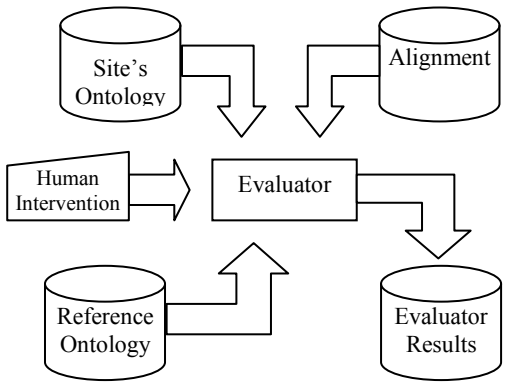


**Figura 5. Estructura del Evaluator**

El Evaluator emplea la salida producida por el Matcher e intenta evaluar las diferencias existentes, no solo a nivel de las estructuras individualmente sino globalmente entre ambas ontologías. Este proceso requiere también de la intervención humana.

El proceso de evaluación implicará medir la correspondencia entre ambas ontologías para lo cual el Evaluator deberá emplear distintas métricas. Es aquí donde se deberán analizar diferentes métricas o medidas de similitud que permitan establecer el grado de similitud entre las ontologías.

Dado que las ontologías de gobierno electrónico no son ampliamente usadas en la actualidad y que la gran mayoría de los sitios de e-gov emplean tecnologías HTML o XML es que, también estamos considerando emplear técnicas de extracción de conocimiento a partir de documentos no estructurados (Text Mining, Web Content Mining y Web Structure Mining).

# Líneas de Investigación y Desarrollo

Este trabajo es llevado a cabo dentro de la línea de "Métodos Formales y Prototipos Evolutivos" del proyecto de incentivos de la Universidad Nacional de San Luis, código 22/F822: "Ingeniería de Software: Conceptos, Métodos y Herramientas en un Contexto de Ingeniería de Software en Evolución" y se encuentra íntimamente relacionado con

trabajos previos, publicados por los autores [CDF09], [DDF07], en el área del desarrollo de modelos de evaluación de sitios de gobierno electrónico.

Cabe destacar que en esa misma línea de investigación, hemos también aplicado el método LSP para otro tipo de sistemas en trabajos tales como [DDF00], [DFP04], [DFPS01], [DPS03], [FDD00], [FDPS05], [MDU00]. Asimismo, en este contexto, como parte de un trabajo de tesis de maestría [Cas10], hemos realizado la evaluación de sitios de gobierno electrónico en forma manual y es precisamente a partir del trabajo tedioso de la recolección manual de los datos que surgió la idea de llevar adelante esta nueva línea de investigación. Esperamos que la misma siente las bases para el desarrollo de nuevas tesis de grado y posgrado.

# Formación de Recursos Humanos

La evaluación de sistemas, métodos y herramientas es una de las áreas en la cual hemos estado trabajando desde hace varios años y que ha producido varias publicaciones [DDF00], [CDF09], [DDF07], [FDPS05]. Este trabajo continuo nos ha conducido a la evaluación de sitios de gobierno electrónico, lo que ha dado como resultado una tesis de maestría en 2010, mientras que hay otras en preparación.

Los aspectos propios del trabajo aquí presentado son ambiciosos y se espera que las distintas tareas a desarrollar sirvan para la realización de tesis de posgrado así como de grado.

# Referencias


[CDF09] M. Castro, A. Dasso, A. Funes. "Modelo de Evaluación para Sitios de Gobierno Electrónico". 38 JAIIO/SIE 2009, Simposio de Informática en el Estado 2009, Mar del Plata, Argentina, August 26-28, 2009.

[DB97] J. J. Dujmovic and A. Bayucan, "Evaluation and Comparison of Windowed environments", Proceedings of the IASTED Interna Conference Software Engineering (SE'97), pp 102-105, 1997.

[DDF00] N. Debnath, A. Dasso, A. Funes, G. Montejano, D. Riesco, R. Uzal, "The LSP Method Applied to Human Resources Evaluation and Selection", Journal of Computer Science and Information Management, Publication of the Association of Management/International Association of Management, Volume 3, Number 2, 2000, ISBN 1525-4372, pp.1-12.

[DDF07] Narayan Debnath, Aristides Dasso, Ana Funes, Roberto Uzal, José Paganini. "E-government Services Offerings Evaluation Using Continuous Logic". 2007 ACS/IEEE International Conference on Computer Systems and Applications, AICCSA '2007, Amman, Jordan. Sponsored by IEEE Computer Society, Arab Computer Society, and Philadelphia University, Jordan. May 13-16, 2007

[DFP04] A. Dasso, A. Funes, M. Peralta, C. Salgado, "User Oriented Evaluation Models for DBMSs", 33 Jaiio (ASIS 04), Córdoba, Argentina, 20-24 de Septiembre, 2004.

[DFPS01] A. Dasso, A. Funes, M. Peralta, C. Salgado, "Una Herramienta para la Evaluación de Sistemas", Workshop de Investigadores en Ciencias de la Computación, WICC 2001, Universidad Nacional de San Luis, San Luis, Argentina, May 2001.

[DGC04] Domingue, J., Gutierrez, L., Cabral, L., Rowlatt, M., Davies, R., Galizia, S., WP 9: Case Study eGovernment, D9.3 e-Government ontology. Data, Information and Process Integration with Semantic Web Services, FP6 – 507483. December 14, 2004. Veáse también "D9.3 Annex document: e-Government Ontology (OCML)".

[DPS03] N. Debnath, M. Peralta, C. Salgado, A. Funes, A. Dasso, D. Riesco, G. Montejano, R. Uzal, "Web Programming

Language Evaluation using LSP", CAINE03 Proceedings, Las Vegas, USA, 11-13 de Noviembre, 2003. ISBN: 1-880843-49-8, pp 302-305.

[Duj07] Jozo J. Dujmovic, "Continuous Preference Logic for System Evaluation", IEEE Transactions on Fuzzy Systems, Vol. 15, Nº 6, December 2007.

[Duj96] J. J. Dujmovic, "A Method for Evaluation and Selection of Complex Hardware and Software Systems", The 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems. CMG96 Proceedings, vol. 1, pp.368-378, 1996.

[Duj97] J. J. Dujmovic, "Quantitative Evaluation of Software", Proceedings of the IASTED International Conference on Software Engineering, edited by M.H. Hamza, pp. 3-7, IASTED/Acta Press, 1997.

[FDD00] A. Funes, A. Dasso, J. Dujmovic, G. Montejano, D. Riesco, R. Uzal, "Web Browsers Performance Analysis using LSP Method", Proceedings de la International Conference on Software Engineering Applied to Networking & Parallel/Distributed Computing (SNPD'00), Mayo, 2000, Reims, Francia. ISBN: 0-9700776-0-2, pp 551-558.

[FDPS05] Ana Funes, Aristides Dasso, Carlos Salgado, Mario Peralta, "UML Tool Evaluation Requirements". Argentine Symposium on Information Systems ASIS 2005. Rosario, Argentina. September 29-30, 2005.

[LHL01] Berners-Lee, Tim; James Hendler and Ora Lassila, The Semantic Web. Scientific American Magazine. May 17, 2001. http://www.sciam.com/article.cfm?id=the-semantic-web&print=true. Retrieved March 26, 2008.

[MDU00] G. Montejano, J.J. Dujmovic, R. Uzal, D. Riesco, A. Dasso, A. Funes, "A Prototype Tool for Decision Support based in the LSP Method", Proceedings de IASTED, Las Vegas, Nevada, USA, 6-9 de Noviembre, 2000. ISBN: 0-88986-306-7, pp 1-4.

[Cas10] Castro, Marcelo; "Análisis de las propiedades y atributos propios de sitios de gobierno electrónico", Tesis para la Maestría en Ingeniería del Software. Departamento de Informática, Universidad Nacional de San Luis, 2010.

[EC02] European Commission benchmarks for on-line availability of public services: – "Common list of basic public services". http://ec.europa.eu/information_society/eeurope/2002/action_plan/pdf/basicpublicservices.pdf. Retrieved 9/21/2006.